Cross-Modal-Domain Generalization Through Semantically Aligned Discrete Representations

Souptik Sen Raneen Younis Zahra Ahmadi

Peter L. Reichertz Institute for Medical Informatics,
Hannover Medical School, Germany

{Sen.Souptik, Younis.Raneen, Ahmadi.Zahra}@mh-hannover.de

Abstract

*Multimodal learning seeks to integrate information across diverse sensory sources, yet current approaches struggle to balance cross-modal generalizability with modality-specific structure. Continuous (implicit) methods preserve fine-grained priors but render generalization challenging, while discrete (explicit) approaches enforce shared prototypes at the expense of modality specificity. We introduce **CoDAAR**[1] (**C**ross-m**o**dal **D**iscrete **A**lignment **A**nd **R**econstruction), a novel framework that resolves this long-standing trade-off by establishing semantic consensus across modality-specific codebooks through index-level alignment. This design uniquely allows CoDAAR to preserve modality-unique structures while achieving generalizable cross-modal representations within a unified discrete space. CoDAAR combines two complementary mechanisms: Discrete Temporal Alignment (DTA), which enables fine-grained temporal quantization, and Cascading Semantic Alignment (CSA), which promotes progressive cross-modal semantic agreement. Together, they establish a competition-free unified representation space. Trained with self-supervised reconstruction objectives on paired multimodal sequences, CoDAAR demonstrates robust cross-modal and cross-domain generalization. Across Cross-Modal Generalization benchmarks, including event classification, localization, video segmentation, and cross-dataset transfer, CoDAAR achieves state-of-the-art performance, establishing a new paradigm for discrete and generalizable multimodal representation learning.*

1. Introduction

Humans perceive the world through a continuous stream of multimodal sensory inputs: auditory, visual, and linguistic. Yet, our cognitive system encodes these experiences as discrete conceptual units rather than continuous representations. For example, a video of a departing train is abstracted into language tokens like "train", "moving", "station", allowing our mind to decompose multimodal stimuli into symbolic components. This discretized architecture enables cross-modal knowledge transfer; for instance, understanding "train" in text supports immediate recognition in visual or auditory contexts, supporting higher-order reasoning and generalization across modalities.

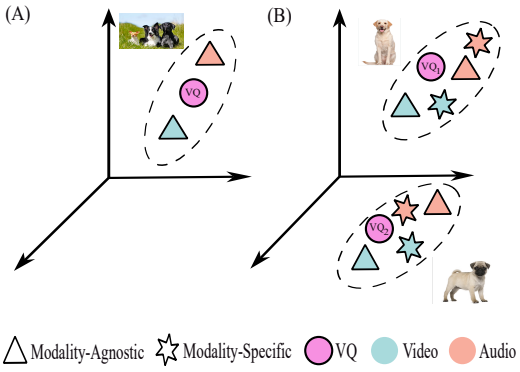


Figure 1. (A) Current SOTA unified multimodal representations map only modality-agnostic features to discrete codes. (B) Our proposed representation maps both (i) modality-agnostic and (ii) modality-specific features within a unified space.

Inspired by this cognitive principle, recent studies have explored unified discrete representations for cross-modal generalization and reasoning. Discrete spaces compress high-dimensional multimodal features into finite sets of generalizable latent codes [22, 34], enabling tasks such as visual question answering [3, 18, 19], query-based segmentation [10, 29], and audio-visual event localization [32, 40]. Existing methods generally follow two paradigms: (i) *Implicit* continuous models [1, 26, 35, 44], which project modalities into distinct continuous embedding spaces and align them contrastively; and (ii) *Explicit* discrete approaches [9, 14, 22, 39, 42], which quantize multimodal features via shared codebooks or prototypes. Implicit

[1]Our code is available at https://github.com/EMuLeMultimodal/CoDAAR

models offer flexibility and preserve modality uniqueness but lack generalizability, while explicit methods often lose modality-specific structure by enforcing a shared generalizable space. Furthermore, most discrete frameworks [9, 42] are coarse-grained, collapsing sequential features into single prototypes before quantization, discarding temporal semantics. As a result, they remain limited to simple retrieval tasks [9, 42] or struggle to retain modality-specific cues needed for complex downstream reasoning [15, 16, 39].

The significance of preserving modality-specific semantics becomes clear in unconstrained real-world scenarios. Distinguishing visually similar entities, such as a pug versus a labrador, requires nuanced visual cues like fur texture, ear shape, or gait, beyond the generic concept of "dog" [19, 29] (Fig. 1). Similarly, in audio-visual data, differentiating a violin from a viola depends on the subtle acoustic timbre and visual morphology of these instruments that coarse semantic alignment alone cannot capture [32, 40]. Fine-grained discrete codebook methods [14–16, 22, 39] that rely on a unified quantization encounter a fundamental challenge: *representation competition*. High-variance modalities, primarily vision with its rich spatio-temporal complexity, dominate the shared codebook, biasing codeword positioning towards visual semantics, simultaneously relegating low-variance modalities (e.g., audio) to suboptimal regions [20]. This disparity results in an unfavourable trade-off: sacrificing modality-specific structural priors for cross-modal generalization, or preserving modal heterogeneity at the cost of generalizability [10, 19, 43]. This raises a central question: *How can we construct multimodal representations that preserve comprehensive modality-specific information while achieving generalizable cross-modal representations?*

To address these challenges, we propose the **C**ross-**Mo**dal **D**iscrete **A**lignment **A**nd **R**econstruction (**CoDAAR**) architecture. CoDAAR introduces modality-specific codebooks instead of a single shared discrete space, directly reducing representational competition. Each modality preserves its intrinsic structural priors within its respective codebook, whereas semantic alignment and generalization arise through index-level correspondence across codebooks. Our architecture comprises two key components: (1) ***Discrete Temporal Alignment*** **(DTA)** module handles fine-grained temporal semantics via cross-modal exponential moving average updates at synchronized timeframes across modalities. Each modality codeword aggregates weighted contributions from both fine-grained self-modal features (dominant) and time-synchronized cross-modal streams (auxiliary), thereby balancing the influence of high-variance modalities. (2) ***Cascading Semantic Alignment*** **(CSA)** module enforces semantic consensus at the codebook index level by hierarchically aligning corresponding codewords across modalities. This module gravitates the modal-specific codewords at the same index towards a common semantic mean, ensuring that identical indices across codebooks represent similar semantics. This design combines the strengths of both implicit and explicit approaches, preserving modality-specific cues while achieving generalizable cross-modal representations. Our main contributions are as follows:

- We introduce **CoDAAR**, a discrete multimodal alignment framework with two novel modules: **DTA** for fine-grained temporal quantization through synchronized cross-modal updates, and **CSA** for hierarchical index-level semantic alignment across modality-specific codebooks.
- Our extensive experiments demonstrate cross-modal generalization under scarce annotation and variable labeling costs across modalities, on the Cross-Modal Generalization (CMG) benchmarks [39]. This evaluation assesses CoDAAR's capacity for zero-shot transfer between labeled and unlabeled modalities. CoDAAR achieves state-of-the-art performance across diverse downstream tasks involving unseen modalities.

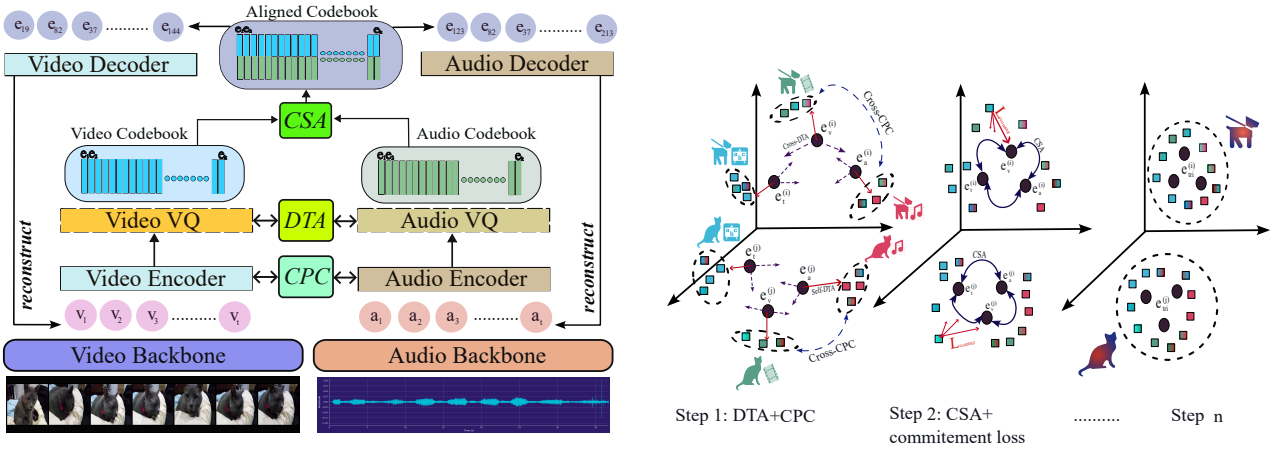
2. Related Work

Implicit Multimodal Representations

Implicit multimodal methods align modalities within continuous embedding spaces via contrastive learning [26, 35], spanning vision-language [6, 23, 26, 37, 41], audio-visual [12, 17, 24, 44], video-audio-text [1, 27, 38], and speech-text [4, 11, 36] combinations through contrastive objectives, modality-agnostic transformers, and cross-modal knowledge distillation [2, 25, 28]. While achieving strong cross-modal semantic consistency, their unbounded continuous embeddings inherently restrict generalizability and lack explicit structural similarity across modalities, inhibiting cross-modal generalization.

Explicit Multimodal Representations

Explicit methods achieve generalization through discrete quantization using shared codebooks [22, 34] or prototypes [9] to enable explicit cross-modal similarity. Coarse-grained approaches [9, 42] use Optimal Transport or self-cross-reconstruction for prototype alignment but lose temporal semantics by condensing sequences into single vectors, restricting them to basic retrieval tasks. Recent fine-grained methods [14–16, 22, 39] preserve temporal structure through frame-level quantization, improving performance on complex applications [16, 19, 43]. However, their unified codebook designs create fundamental limitations: high-variance modalities dominate codeword optimization while modality-specific structural priors diminish [15, 16, 39]. Our approach addresses these constraints using modality-specific codebooks that preserve structural



(a) Model architecture.

(b) Cross-modal Discrete Alignment

Figure 2. The overview of our proposed CoDAAR framework. (a) Model architecture. (b) The Cross-modal Discrete Alignment mechanism, comprising Discrete Temporal Alignment (DTA), CPC loss, Cascading Semantic Alignment (CSA), and commitment loss.

priors while achieving semantic consistency through index-level alignment across codebooks.

3. Cross-Modal Generalization Task Definition

The *Cross-Modal Generalization* (CMG) task was introduced in [39] to assess whether multimodal models can learn modality-invariant discrete representations. This task measures how supervision obtained from one modality transfers to another during downstream evaluation. Let $m_1, m_2 \in \{a, v, t\}$ denote audio, video, and text modalities with $m_1 \neq m_2$. During downstream training, the model learns representations for input samples $\mathbf{x}_i^{m_1}$ and their respective labels $\mathbf{y}_i^{m_1}$, using a modality-specific encoder $\Psi^{m_1}(\cdot)$ and a modality-invariant task-head decoder $\mathcal{G}(\cdot)$. The encoder produces continuous embeddings, which are then mapped to discrete latent codes using a vector-quantization operator $VQ(\cdot)$. The decoder is trained on top of these codes using an evaluation loss $\mathcal{E}_{\mathcal{L}}(\cdot)$. At test time, the decoder is evaluated on samples $\mathbf{x}_i^{m_2}$ and labels $\mathbf{y}_i^{m_2}$ from a different modality m_2. The parameters of both Ψ^{m_1} and Ψ^{m_2} stay static throughout training and testing, while only the parameters of $\mathcal{G}(\cdot)$ are updated during training. Performance reflects the extent to which the learned discrete space aligns heterogeneous modalities and supports zero-shot knowledge transfer from m_1 to m_2. A complete notation library is provided in the supplementary material.

4. CoDAAR Architecture

4.1. Cross-Modal Discrete Alignment Framework

We introduce **C**ross-**Mo**dal **D**iscrete **A**lignment **A**nd **R**econstruction (CoDAAR), a framework that constructs a unified discrete latent space for fine-grained cross-modal and cross-domain generalization. Given a paired multimodal dataset $\mathbf{X} = \{(x_i^a, x_i^v, x_i^t)\}_{i=1}^N$ with aligned audio, video, and text sequences, CoDAAR assigns each modality m a codebook $\mathbf{E}_m \in \mathbb{R}^{K \times D}$ containing K codewords. Each modality's input is encoded into continuous semantic embeddings, then discretized via these codebook indices. The key innovation is aligning these indices such that identical indices across modalities represent the same latent concept, yielding a unified discrete vocabulary $\mathbf{E} = [\mathbf{E}_a; \mathbf{E}_v; \mathbf{E}_t]$. This aligned index space enables cross-modal knowledge transfer while preserving modality-specific structural cues in respective codebooks through two complementary objectives: reconstruction for capturing modality-specific details and alignment for enforcing cross-modal semantic coherence, supporting evaluation under the CMG protocol (Section 3). Domain generalization emerges through unsupervised pre-training on these structural and semantic patterns. Figure 2a illustrates the architecture.

4.1.1. VQ-VAE Backbone

Given paired multimodal sequences $\mathbf{X} = \{(x_i^a, x_i^v, x_i^t)\}_{i=1}^N$, the core backbone module follows a vector-quantized reconstruction pipeline. For each modality $m \in \{a, v, t\}$, an encoder $\Psi^m(\cdot)$ maps inputs x_i^m to a continuous latent representation $z_i^m = \Psi^m(x_i^m) \in \mathbb{R}^{T \times D}$, where T and D denote the temporal length and feature dimension. Each modality maintains a codebook $\mathbf{E}_m \in \mathbb{R}^{K \times D}$ with K codewords of the same dimensionality as the latent embeddings. Encoder embeddings are discretized to these codebooks via a nearest-neighbour lookup. Each vector frame $z_{i,t}^m, t \in \{1{:}T\}$, is mapped to a quantized embedding, $\hat{z}_{i,t}^m = \text{VQ}(z_{i,t}^m) = \mathbf{e}_m(k)$, where the index k is selected by $k = \arg\min_j \|z_{i,t}^m - \mathbf{e}_m(j)\|_2^2$. For reconstruction, the quantized embeddings $\hat{z}_i^m$ are

concatenated with a modality-specific projection $P_m(x_i^m)$ of the input features, and passed through a decoder $\mathcal{D}_m(\cdot)$ to obtain $\tilde{x}_i^m = \mathcal{D}_m([\hat{z}_i^m; P_m(x_i^m)])$. Training follows the standard VQ-VAE objective [34], combining reconstruction, quantization, and commitment terms:

$$\mathcal{L} = \underbrace{\|x_i^m - \tilde{x}_i^m\|_2^2}_{\text{reconstruction loss}} + \underbrace{\|\text{sg}[z_i^m] - \mathbf{e}_m\|_2^2}_{\text{VQ loss}} + \underbrace{\beta\|z_i^m - \text{sg}[\mathbf{e}_m]\|_2^2}_{\text{commitment loss}}, \tag{1}$$

where $\text{sg}[\cdot]$ is the stop-gradient operator and $\beta = 0.25$ in all experiments [34]. Following [34, 42], we construct the codebook via Exponential Moving Average (EMA) operation instead of the explicit VQ loss, improving stability and preventing codebook collapse. Without additional constraints, this core formulation learns modality-specific codebooks that remain semantically misaligned across modalities. We use this backbone to isolate the effects of our alignment mechanism, ensuring that improvements arise solely from cross-modal correspondence.

4.1.2. CPC-based Cross-Modal Information Maximization

To inject multimodal semantics into our unimodal encoders $\Psi^m(\cdot)$, we adopt a Contrastive Predictive Coding (CPC) objective [35, 39] that makes each encoder cross-modally aware. For a modality pair (a, b), a unidirectional LSTM summarizes past embeddings of modality a into a context vector c_t^a, which is optimized to predict future embeddings of modality b. This incorporates fine-grained temporal cross-modal cues directly into the unconstrained unimodal streams, improving modal transferability during downstream inference. We apply the objective symmetrically across ordered modality pairs and average their contributions:

$$\mathcal{L}_{\text{CPC}} = -\frac{1}{H}\sum_{h=1}^{H} \log \frac{\exp\big((z_{t+h}^b)^\top W_h^a\, c_t^a\big)}{\sum_j \exp\big((z_j^b)^\top W_h^a\, c_t^a\big)}, \tag{2}$$

where c_t^a is the modality-a context at time t; z_{t+h}^b is the future latent of modality b; W_h^a is a step-specific linear projection; the denominator aggregates one positive and sampled negatives $\{z_j^b\}$ from modality b.

4.1.3. Cross-Modal Discrete Alignment

With our CPC-induced cross-modally aware encoders $\Psi^m(\cdot)$ in place, we introduce a three-fold alignment to learn our codebooks. *DQA* aligns codebook index selection across modalities; *DTA* temporally aligns the construction of these selected codewords via cross-modal EMA; and *CSA* applies a cascading shift to these constructed codewords to enforce cross-index semantic alignment.

Discrete Quantization Alignment (DQA)

To ensure index utilization of an embedding i is consistent across modalities $m \in \{a, v, t\}$, we employ the Cross-Modal Code Matching (CMCM) [22] loss, extending it to our modality-specific codebooks $\{\mathbf{E}_m\}_{m\in\{a,v,t\}}$. For each paired embedding (z_i^a, z_i^v, z_i^t), we compute the sequence-level code-usage distributions p_m^i over indices $k \in \{1, \ldots, K\}$ and align these distributions across modalities. We match the embedding distributions of corresponding indices by contrasting distributions from modality-paired sample embeddings while treating non-matching samples in the batch as negatives. This encourages the model to assign the same index to the same embedding instance across modalities, leading to consistent nearest-neighbour lookups even though each modality quantizes through a different codebook. DQA therefore promotes cross-modal index usage synchronization. However, it does not guarantee semantically aligned codebook constructions. The DQA loss equation can be found in the supplementary material.

Discrete Temporal Alignment (DTA)

We introduce DTA to enable fine-grained temporal quantization and to *align* how selected codewords are *constructed* over training iterations. We treat codewords as k-means centroids for the latent encoder embeddings and update them using a temporally aligned, cross-modal EMA. For each timestep of a paired sample embedding, the centroid of modality m moves toward a convex combination of its own quantized embeddings and the paired embeddings from the other modalities at that same timestep. Cross-modally aware encoders produce temporally aligned embeddings with comparable semantics at identical timesteps within a sample. This allows our cross-modal EMA updates to aggregate the underlying aligned semantics.

Indexing: A mini-batch contains B samples indexed by $i \in \{1{:}B\}$, each with T timesteps. For modality $m \in \{a, v, t\}$, the latent encoder embedding is $\mathbf{z}_i^m \in \mathbb{R}^{T\times D}$.

Hard assignment (per codeword k): For modality m with codebook $\mathbf{E}_m = \{\mathbf{e}_m(k)\}_{k=1}^K$, each embedding frame $\mathbf{z}_{i,t}^m, t \in \{1{:}T\}$, selects exactly one codeword via nearest-neighbour lookup:

$$\begin{aligned} \pi_m(i,t) &= \arg\min_{j\in\{1:K\}} \left\|\mathbf{z}_{i,t}^m - \mathbf{e}_m(j)\right\|_2^2, \\ q_{m,i,t}(k) &= \mathbb{1}[k = \pi_m(i,t)]\,. \end{aligned} \tag{3}$$

Stacking $q_{m,i,t}(k)$ over (i,t) produces a binary responsibility tensor with one-hot rows per (i,t).

Temporally aligned accumulators: For $\{n, p\} = \{a, v, t\} \setminus \{m\}$, we define EMA inputs that aggregate matched (i,t) assignments:

$$\begin{aligned} \mathbf{U}_m^{\text{self}}(k) &= \sum_{i=1}^{B}\sum_{t=1}^{T} q_{m,i,t}(k)\, \mathbf{z}_{i,t}^m, \\ \mathbf{U}_m^{\text{cross}}(k) &= \sum_{i=1}^{B}\sum_{t=1}^{T} q_{m,i,t}(k)\, \big(\mathbf{z}_{i,t}^n + \mathbf{z}_{i,t}^p\big). \end{aligned} \tag{4}$$

EMA update (per codeword k): We use λ_{self}=0.6, λ_{cross}=0.2 (two cross terms; 0.6+0.2+0.2 = 1), decay $\rho \in (0,1)$, and $\varepsilon > 0$. The update is:

$$\begin{aligned}
\mathbf{H}_m(k) &= \lambda_{\text{self}} \mathbf{U}_m^{\text{self}}(k) + \lambda_{\text{cross}} \mathbf{U}_m^{\text{cross}}(k), \\
\mathbf{S}_m^{\text{new}}(k) &= \rho\, \mathbf{S}_m(k) + (1-\rho)\, \mathbf{H}_m(k), \\
C_m^{\text{new}}(k) &= \rho\, C_m(k) + (1-\rho) \sum_{i=1}^{B} \sum_{t=1}^{T} q_{m,i,t}(k), \\
\mathbf{e}_m^{\text{new}}(k) &= \frac{\mathbf{S}_m^{\text{new}}(k)}{C_m^{\text{new}}(k) + \varepsilon}.
\end{aligned} \tag{5}$$

Here, $\mathbf{S}_m(k)$ tracks the EMA-weighted sum of assigned features (self and time-aligned cross-modal), $C_m(k)$ tracks the EMA-weighted assignment count, and $\mathbf{e}_m^{\text{new}}(k)$ is the resulting centroid. The weights $(0.6, 0.2, 0.2)$ form a convex partition. The self-term anchors each codeword in its own modality. The cross-terms incorporate time-aligned evidence from the paired modalities at the same (i,t), acting as soft teachers. Updating the codebooks separately for each modality with dominant self-anchors, unlike methods with a shared codebook [16, 39], prevents competition in a single shared space and avoids dominance by any one modality.

Cascading Semantic Alignment (CSA)

Although the unimodal codebooks already mix modality-specific and shared cues through DTA, index-level semantic misalignment arises from independent initialization across $\mathbf{E}_a$, $\mathbf{E}_v$, and $\mathbf{E}_t$ and modality biases. To enforce semantic consistency across modality-specific codebooks, we introduce this hierarchical cascade module that gravitates the three centroids at index k toward a shared multimodal semantic mean anchored by $\mathbf{c}^0(k)$ (Eq. 6). The cascade mitigates the misalignment by applying a sequential T→A→V centroid shift (Fig. 3), ordered by increasing relative semantic capacity: text (global semantics) < audio (temporal semantics) < video (spatio-temporal semantics). Text is first pulled toward the cross-modal anchor $\mathbf{c}^0(k)$; audio then bridges text and video; video finalizes the consensus while retaining the most detail (Fig. 3). Intuitively, the multimodal semantic mean tends to lie closer to the higher-capacity stream, i.e., video, which contributes additional spatial detail. Let $\mathbf{e}_v^0(k), \mathbf{e}_a^0(k), \mathbf{e}_t^0(k) \in \mathbb{R}^D$ be post-DTA centroids at index k in a mini-batch iteration. We define $\mathbf{c}^0(k)$, the geometric centroid of the 3 codewords, as the cross-modal semantic anchor for index k.

$$\mathbf{c}^0(k) = \tfrac{1}{3}\big(\mathbf{e}_v^0(k) + \mathbf{e}_a^0(k) + \mathbf{e}_t^0(k)\big). \tag{6}$$

We apply cascading equal-weight updates to the modality centroids:

$$\begin{aligned}
\mathbf{e}_t^1(k) &= \mathbf{c}^0(k), \\
\mathbf{e}_a^1(k) &= \tfrac{1}{3}\big(\mathbf{e}_a^0(k) + \mathbf{e}_v^0(k) + \mathbf{e}_t^1(k)\big), \\
\mathbf{e}_v^1(k) &= \tfrac{1}{3}\big(\mathbf{e}_v^0(k) + \mathbf{e}_a^1(k) + \mathbf{e}_t^1(k)\big).
\end{aligned} \tag{7}$$

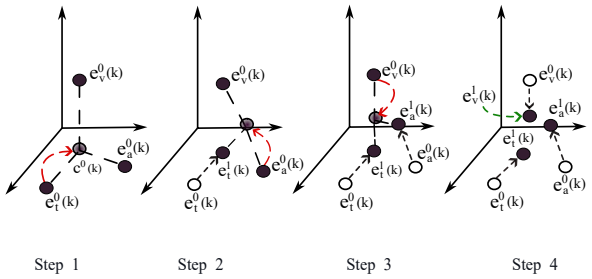


Figure 3. Cascading Semantic Alignment Visualized

Thus, the modality-specific centroids at index k shift toward distinct multimodal locations in the representation space rather than collapsing into a single trimodal point. This allows the centroids to share a unified semantic meaning while retaining modality-specific characteristics, with video (last) retaining the most. Learned mixing weights destabilize our EMA update and allow one modality to dominate, causing codebook collapses. Therefore, we use fixed, non-negative update weights in Eq. 7; that sum to 1 and create progressive centroids. This keeps each update inside the convex hull of its contributors and stabilizes codebook construction. As training proceeds, and **DTA** aggregates more paired embeddings, the three codewords at index k move toward semantic consensus in an independent yet coordinated manner through this complementary cascade module. The commitment loss (Eq. 8) then pulls multimodal latents toward their assigned centroids, forming *distinct semantic spheres* in the multimodal representation space (Fig. 2b).

4.1.4. Pretraining Objectives

Cross-modal Commitment loss: For modality $m \in \{a, v, t\}$, let $z_i^m = \Psi^m(x_i^m)$ be the encoder embedding of sample i, $z_{i,t}^m$, $t \in \{1{:}T\}$, the embedding frame, and let $k_{i,t}$ be the index selected by the quantizer (nearest code) in modality m, with codewords $\mathbf{e}_m(k) \in \mathbf{E}_m$. Using the stop–gradient operator $\text{sg}[\cdot]$, we define

$$\begin{aligned}
\mathcal{L}_{\text{commit}}^m = \sum_{t=1}^{T} \Big[& \beta \| z_{i,t}^m - \text{sg}[e_m(k_{i,t})] \|_2^2 \\
& + \frac{\beta}{2} \sum_{n \neq m} \| z_{i,t}^m - \text{sg}[e_n(k_{i,t})] \|_2^2 \Big].
\end{aligned} \tag{8}$$

where $n \in \{a, v, t\}$, and $\beta = 0.25$. The first term makes each encoder embedding *commit* to its own selected codeword; the (weaker) cross terms $(\beta/2)$ softly pull it toward the same-index centroids of the other modalities, encouraging index-level alignment.

Cross-modal reconstruction: After aligning the cross-modal codeword indices, we get $\hat{\mathbf{z}}_i^m \in \mathbb{R}^{T \times D}$ as the quantized latent of modality $m \in \{a, v, t\}$. Thus, we can define the concatenated tri-modal code as: $\hat{\mathbf{z}}_i^{\text{tri}} = [\hat{\mathbf{z}}_i^a; \hat{\mathbf{z}}_i^v; \hat{\mathbf{z}}_i^t] \in$

$\mathbb{R}^{T\times 3D}$. Each modality is reconstructed from the *same* trimodal code, instead of the unimodal codes, concatenated with the modality-specific projection as defined in Sec. 4.1.1:

$$\begin{aligned}\tilde{\mathbf{x}}_i^m &= \mathcal{D}_m\big([\hat{\mathbf{z}}_i^{\text{tri}};\, P_m(\mathbf{x}_i^m)]\big),\\ \mathcal{L}_{\text{rec}}^m &= \big\|\mathbf{x}_i^m - \tilde{\mathbf{x}}_i^m\big\|_2^2.\end{aligned} \tag{9}$$

This introduces cross-modal reconstruction gradients that flow to all unimodal encoders $\Psi^m(\cdot)$, making them cross-modally aware.

Final Loss: Our pre-training minimizes the following composite objective:

$$\begin{aligned}\mathcal{L}_{\text{total}} = &\sum_{m\in\{a,v,t\}} \big(\mathcal{L}_{\text{rec}}^m + \mathcal{L}_{\text{commit}}^m\big)\\ &+ \sum_{(m,n)\in\mathcal{P}} \big(\mathcal{L}_{\text{CPC}}^{m\leftrightarrow n} + \mathcal{L}_{\text{CMCM}}^{m\leftrightarrow n}\big).\end{aligned} \tag{10}$$

Here $\mathcal{P} = \{(a,v),(a,t),(v,t)\}$ denotes modality pairs available in a batch. **Reset Code:** Following [39], we *reset* any inactivated code that is unselected for N_{re} consecutive batches by reinitializing it from an active code plus small noise, preventing dead codes.

4.2. Downstream Protocol

After pretraining, *we freeze all our encoders and codebooks*. Downstream protocols involve training only lightweight task heads with an evaluation loss $\mathcal{E}_{\mathcal{L}}(\cdot)$ following the CMG evaluation setup in Section 3.

5. Experiments

5.1. Evaluation Setup

We evaluate CoDAAR under two pretraining settings: **AV** (Audio–Video) and **AVT** (Audio–Video–Text). Challenging cross-modal and cross-domain transfer experiments use only AVT due to AV's lack of textual semantic grounding. We pretrain on VGGSound-AVEL [5, 44] using 40k and 90k splits, which contain audio, video, and event labels; for AVT, we incorporate text via event-label descriptions from [39].

1. Cross-modal Event Classification (AVE) [32]: Each video contains a single event label. We train a global event classifier using one modality (e.g., video) and evaluate zero-shot performance on another (e.g., audio). Precision is reported as the evaluation metric.

2. Cross-modal Event Localization (AVVP) [33]: This multi-label dataset includes temporally overlapping events. We train a fine-grained event classifier on one modality and test on another, measuring segment-level accuracy to capture temporal alignment and class imbalance.

3. Cross-Modal Zero-Shot Retrieval (MSCOCO [21], Clotho [8]): We evaluate zero-shot cross-modal retrieval on MSCOCO for vision↔text and Clotho for audio↔text by retrieving nearest neighbours in the codebook-quantized embedding space. Recall@k ($k \in \{1,5,10\}$), which measures whether the correct match appears within the top-k retrieved candidates, is reported.

4. Cross-modal Video Segmentation (AVSBench-S4) [43]: Using the AVT setup, we train a query-based video segmenter with one modality (e.g., text) and directly test segmentation performance in another modality (e.g., audio). We report mean Intersection-over-Union (mIoU) and F1 score.

Precision is used for single-label datasets (AVE), while segment-level F1 and accuracy are reported for multi-label or localization tasks (AVVP, AVSBench).

Additional cross-dataset transfer experiments are provided in the supplementary material.

5.2. Implementation Details

We use the unimodal discretization framework from Sec. 4.1.1 as our baseline and additionally compare CoDAAR with state-of-the-art multimodal discrete representation and domain generalization methods: CMCM [22], CODIS [9], TURN [42], DCID [39], and MICU [16]. Since CMCM, CODIS, and TURN primarily align only *two* modalities and cannot reliably generalize to unconstrained tri-modal setups, we compare AVT settings only with DCID and MICU, which support tri-modal alignment. All experiments use embedding dimension D=256 and K=800 codewords per modality. Temporal EMA updates use λ_{self}=0.6, λ_{cross}=0.2 for tri-modal (AVT) and λ_{self}=0.75, λ_{cross}=0.25 for bi-modal (AV). Complete implementation details can be found in the supplementary material.

5.3. Comparison with SOTA

Cross-modal Event Classification and Localization: Tables 1 and 2 report results for the AV and AVT pretraining settings on AVE (global classification) and AVVP (segment-level localization). Without cross-modal alignment, the unimodal baseline transfers knowledge poorly between modalities. Among prior methods, DCID performs well for smaller datasets, while MICU yields stronger classification but weaker localization. CoDAAR achieves the most balanced and consistent results, particularly on AVVP. While DCID and MICU's unified codebooks produce strictly modal-agnostic representations, our discrete representations retain both modality-specific cues and modality-agnostic semantics, enabling fine-grained localized multi-event segment-level decisions. In the larger 90K-scale setting, CoDAAR maintains stable accuracy, whereas DCID and MICU degrade, suggesting that our temporal–hierarchical alignment helps mitigate overfitting from instances with noisy multimodal alignment in larger datasets. Incorporating text during pretraining (AVT) fur-

Method	VGGSound-AVEL 40K				VGGSound-AVEL 90K				Average
	AVE		AVVP		AVE		AVVP		
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	
Baseline	5.1	4.5	5.5	4.3	7.5	10.2	6.2	7.6	6.4
CMCM [22]	32.7	36.8	41.9	45.1	30.5	33.7	38.4	43.9	37.9
CODIS [9]	20.8	26.4	35.1	37.9	29.3	31.1	33.8	36.4	31.4
TURN [42]	19.1	24.3	36.9	39.3	28.5	32.2	32.5	37.6	31.3
DCID [39]	**47.7**	**52.3**	**64.0**	65.6	34.8	49.0	59.7	64.8	54.7
MICU [16]	47.2	51.4	38.4	33.5	33.0	37.7	42.9	46.1	41.3
CoDAAR (ours)	47.6	49.7	63.6	**72.4**	**43.3**	**49.4**	**60.0**	**67.2**	**56.6**

Table 1. AV setting: comparison with state-of-the-art methods on (i) event classification (AVE, Precision) and (ii) event localization (AVVP, Segment-level Accuracy). V→A / A→V denote transfer directions.

Method	VGGSound-AVEL 40K				VGGSound-AVEL 90K				Average
	AVE		AVVP		AVE		AVVP		
	V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V	
Baseline	6.2	5.7	7.1	6.5	8.3	9.1	8.6	9.4	7.6
DCID [39]	54.1	55.0	63.4	71.0	43.7	50.3	64.0	64.2	58.2
MICU [16]	**56.1**	**57.1**	59.5	56.2	43.1	47.8	45.6	47.1	51.6
CoDAAR (ours)	52.3	55.5	**70.8**	**72.5**	**50.8**	**51.9**	**69.7**	**70.4**	**61.7**

Table 2. AVT setting: comparison with state-of-the-art methods on (i) event classification (AVE, Precision) and (ii) event localization (AVVP, Segment-level Accuracy).

ther improves cross-modal alignment. Text event descriptions act as holistic semantic anchors, stabilizing codebook index synchronization. While MICU and DCID achieve slightly higher AVE precision in small-scale setups, CoDAAR outperforms on the multi-event AVVP benchmark and remains robust across scales.

Cross-Modal Zero-Shot Retrieval: Table 3 reports zero-shot retrieval on MSCOCO (V↔T) and Clotho (A↔T). CoDAAR achieves the highest overall average across both scales, with strong gains on Clotho under both 40K and 90K settings, while remaining competitive on MSCOCO at lower thresholds and the 40K setting and leading at R@10 and under 90K. This confirms generalization to unseen datasets and modality pairs without fine-tuning.

Cross-modal Video Segmentation: Table 4 presents results on the AVSBench-S4 benchmark. CoDAAR attains the best or comparable mIoU and F1 scores in both directions (audio ↔ text). The performance advantage is attributed to the tri-modal centroids, which retain richer visual semantics, thereby improving segmentation even when queries are issued from a different modality. Qualitative maps (Figs. 4 and 5) further confirm precise visual localization, demonstrating CoDAAR's ability to generalize to unseen modality combinations. Additional qualitative visualizations can be found in the supplementary material.

Overall, CoDAAR provides consistent improvements across tasks and datasets, resulting in stable cross-modal and cross-domain generalization.

5.4. Ablation Studies

All ablation experiments are conducted on the VGGSound-AVEL 40K split to ensure consistency across analyses.

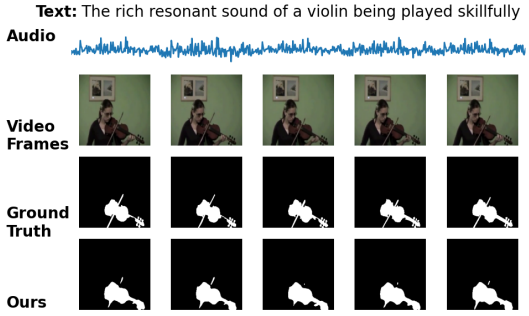


Figure 4. Visualization of audio-to-text generalization on AVS-S4 video segmentation task

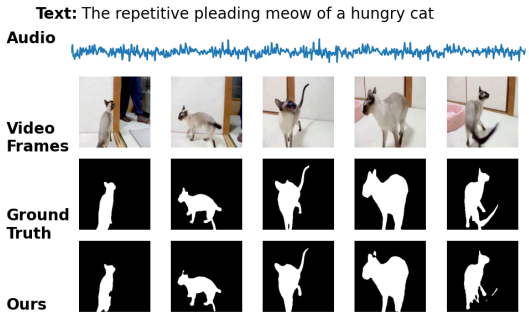


Figure 5. Visualization of text-to-audio generalization on AVS-S4 video segmentation task.

Component Analysis: Tables 5 and 6 report the impact of removing each module under AV and AVT settings. Removing *CSA* leads to the most severe performance degradation (e.g., AVE 21.2/18.1, AVVP 10.4/11.2 in AV), confirming that cross-modal semantic alignment of modality-specific codebooks is essential for index consensus. Elimi-

Method	VGGSound-AVEL 40K						VGGSound-AVEL 90K						Average
	MSCOCO(V↔T)			Clotho(A↔T)			MSCOCO(V↔T)			Clotho(A↔T)			
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
DCID [39]	0.80	**5.00**	8.30	2.06	9.00	16.70	0.80	4.80	8.20	2.77	11.00	20.43	7.49
MICU [16]	**1.30**	**5.00**	8.80	2.44	10.96	18.95	0.90	4.90	9.50	3.35	10.07	17.62	7.82
CoDAAR (ours)	0.80	4.40	**9.50**	**3.30**	**11.57**	**19.00**	**1.30**	**6.10**	**10.40**	**5.64**	**16.65**	**24.70**	**9.45**

Table 3. Cross-modal zero-shot retrieval comparison with discrete SOTA methods on MSCOCO (V↔T) and Clotho (A↔T) datasets.

Methods	VGGSound-AVEL 40K				VGGSound-AVEL 90K				Average
	A→T		T→A		A→T		T→A		
	mIoU	F-score	mIoU	F-score	mIoU	F-score	mIoU	F-score	
DCID [39]	73.3	83.3	**77.7**	**86.7**	74.8	84.7	**76.2**	**86.3**	80.4
MICU [16]	73.8	84.1	76.9	86.5	76.4	86.0	75.7	85.7	80.6
CoDAAR (ours)	**74.3**	**84.5**	75.4	86.6	**76.5**	**86.4**	76.0	86.0	**80.7**

Table 4. AVT setting: comparison with state-of-the-art methods on **AVSBench-S4** query-based video segmentation task, fine-tuned till the 4th downstream epoch. Columns report mIoU and F-score for audio→text and text→audio.

Components				AVE		AVVP	
CPC	DTA	CSA	DQA	V→A	A→V	V→A	A→V
–	✓	✓	✓	38.6	40.2	52.5	53.1
✓	–	–	✓	19.6	9.7	7.3	8.3
✓	✓	–	✓	21.2	18.1	10.4	11.2
✓	–	✓	✓	45.2	47.3	60.5	69.0
✓	✓	✓	–	43.5	49.5	65.9	68.1
✓	✓	✓	✓	**47.6**	**49.7**	**63.6**	**72.4**
Full model							

Table 5. Component Ablation (AV settings): AVE (Precision), AVVP (Segment Accuracy)

Components				AVE		AVVP	
CPC	DTA	CSA	DQA	V→A	A→V	V→A	A→V
–	✓	✓	✓	42.6	42.9	48.1	56.7
✓	–	–	✓	3.4	6.9	4.8	10.8
✓	✓	–	✓	11.1	25.5	10.7	7.5
✓	–	✓	✓	46.9	51.5	66.3	70.6
✓	✓	✓	–	51.5	54.8	69.5	71.0
✓	✓	✓	✓	**52.3**	**55.5**	**70.8**	**72.5**
Full model							

Table 6. Component Ablation (AVT settings): AVE (Precision), AVVP (Segment Accuracy)

Setting	Cascade	AVE		AVVP	
		V→A	A→V	V→A	A→V
AV	V→A	46.5	**50.0**	62.6	70.3
	A→V	**47.6**	49.7	**63.6**	**72.4**
AVT	T→A→V	**52.3**	**55.5**	**70.8**	72.5
	T→V→A	51.4	54.1	69.5	71.0
	A→T→V	51.6	53.6	68.5	71.9
	A→V→T	51.2	52.3	70.5	**73.0**
	V→T→A	51.4	53.8	70.0	71.4
	V→A→T	50.5	52.2	70.1	72.3

Table 7. Cascading order ablation on AVE (Precision) and AVVP (Segment-level Accuracy). AV uses VGGSound-AVEL 40K; AVT exhaustively ablates all six permutations.

nating *CPC* also causes a major drop (AVVP $63.6/72.4 \rightarrow 52.5/53.1$ in AV; $70.8/72.5 \rightarrow 48.1/56.7$ in AVT), as this module injects cross-modal cues into unimodal encoders. The absence of *DTA* or *DQA* individually weakens performance—particularly classification precision when DQA is removed ($52.3/55.5 \rightarrow 51.5/54.8$ in AVT)—indicating that temporal and quantization alignment are complementary. Overall, the full model achieves the best performance in all transfer directions, demonstrating the importance of their joint interaction.

Cascade Order: Table 7 ablates the hierarchical consensus order. In AV, A→V outperforms V→A, suggesting that mapping semantics closer to video codewords benefits from video's richer spatial detail. For AVT, we exhaustively evaluate all six permutations: T→A→V is optimal, as text first provides holistic semantic anchoring while video last refines alignment with spatial cues. Orderings placing text last yield lower AVE scores, confirming its role as an early semantic anchor.

Further ablations on codebook size and embedding dimension are included in the supplementary material.

6. Conclusion

In this paper, we presented **CoDAAR**, a cross-modal discrete alignment module that introduces two novel mechanisms—DTA and CSA—to synchronize modality-specific codebooks at the index level semantically. These modules jointly preserve modality-specific and modality-agnostic information within a unified discrete space. Future work includes expanding CoDAAR to new modalities (e.g., sensor streams, point clouds) and applications such as sentiment analysis and cross-modal retrieval.

7. Acknowledgements

This work was supported by the Federal Ministry of Research, Technology and Space of Germany [project name: EMuLE – Enhancing Data and Model Efficiency in Multimodal Learning; grant number 16IS24059]. The last author was partially funded by the Lower Saxony Ministry of Science and Culture (MWK) with funds from the

Volkswagen Foundation's Zukunft Niedersachsen program [project name: CAIMed - Lower Saxony Center for Artificial Intelligence and Causal Methods in Medicine; grant number: ZN4257]. The authors acknowledge the Hannover Medical School for providing MHH-HPC resources that have contributed to the results reported in this paper.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2

[2] Alex Andonian, Shixing Chen, and Raffay Hamid. Robust cross-modal representation learning with progressive self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16430–16441, 2022. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, 2015. 1

[4] Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing, 2022. 2

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. *CoRR*, abs/2004.14368, 2020. 6

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. {UNITER}: Learning {un}iversal image-{te}xt representations, 2020. 2

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 2

[8] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740, 2020. 6, 3

[9] Jiali Duan, Liqun Chen, Son Tran, Jinyu Yang, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Multi-modal alignment using representation codebook, 2022. 1, 2, 6, 7

[10] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees G. M. Snoek. Actor and action video segmentation from a sentence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5958–5966, 2018. 1, 2

[11] Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. Learning shared semantic space for speech-to-text translation. *CoRR*, abs/2105.03095, 2021. 2

[12] David Harwath, Adrià Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James R. Glass. Jointly discovering visual objects and spoken words from raw sensory input. *CoRR*, abs/1804.01452, 2018. 2

[13] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. Cnn architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2016. 2

[14] Hai Huang, Shulei Wang, and Yan Xia. Semantic residual for multimodal unified discrete representation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. 1, 2

[15] Hai Huang, Yan Xia, Shengpeng Ji, Shulei Wang, Hanting Wang, Minghui Fang, Jieming Zhu, Zhenhua Dong, Sashuai Zhou, and Zhou Zhao. Enhancing multimodal unified representations for cross modal generalization, 2025. 2

[16] Hai Huang, Yan Xia, Shulei Wang, Hanting Wang, Minghui Fang, Shengpeng Ji, Sashuai Zhou, Tao Jin, and Zhou Zhao. Open-set cross modal generalization via multimodal unified representation, 2025. 2, 5, 6, 7, 8, 4

[17] Simon Jenni, Alexander Black, and John Collomosse. Audio-visual contrastive learning with temporal self-supervision, 2023. 2

[18] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1369–1379. Association for Computational Linguistics, 2018. 1

[19] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19108–19118, 2022. 1, 2

[20] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755, 2014. 6, 3

[22] Alexander H. Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James R. Glass. Cross-modal discrete representation learning. *CoRR*, abs/2106.05438, 2021. 1, 2, 4, 6, 7

[23] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of CLIP for end to end video clip retrieval. *CoRR*, abs/2104.08860, 2021. 2

[24] Fabrizio Pedersoli, Dryden Wiebe, Amin Banitalebi, Yong Zhang, George Tzanetakis, and Kwang Moo Yi. Estimating visual information from audio through manifold learning, 2022. 2

[25] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech

recognition with A hybrid ctc/attention architecture. *CoRR*, abs/1810.00108, 2018. 2

[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*. PMLR, 2021. 1, 2

[27] Adrià Recasens, Jason Lin, João Carreira, Drew Jaegle, Luyu Wang, Jean baptiste Alayrac, Pauline Luc, Antoine Miech, Lucas Smaira, Ross Hemsley, and Andrew Zisserman. Zorro: the masked multimodal transformer, 2023. 2

[28] Pritam Sarkar and Ali Etemad. Xkd: Cross-modal knowledge distillation with domain alignment for video representation learning, 2023. 2

[29] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision – ECCV 2020*, pages 208–223. Springer, 2020. 1, 2

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1

[31] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 3, 4

[32] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 6, 3

[33] Yapeng Tian, Dingzeyu Li, and Chenliang Xu. Unified multisensory perception: Weakly-supervised audio-visual video parsing. *CoRR*, abs/2007.10558, 2020. 6, 3

[34] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1, 2, 4

[35] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 1, 2, 4

[36] Chengyi Wang, Yu Wu, Yao Qian, Ken'ichi Kumatani, Shujie Liu, Furu Wei, Michael Zeng, and Xuedong Huang. Unispeech: Unified speech representation learning with labeled and unlabeled data. *CoRR*, abs/2101.07597, 2021. 2

[37] Teng Wang, Wenhao Jiang, Zhichao Lu, Feng Zheng, Ran Cheng, Chengguo Yin, and Ping Luo. Vlmixer: Unpaired vision-language pre-training via cross-modal cutmix, 2022. 2

[38] Zehan Wang, Ziang Zhang, Luping Liu, Yang Zhao, Haifeng Huang, Tao Jin, and Zhou Zhao. Extending multi-modal contrastive representations, 2024. 2

[39] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[40] Haoming Xu, Runhao Zeng, Qingyao Wu, Mingkui Tan, and Chuang Gan. Cross-modal relation-aware networks for audio-visual event localization. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, pages 3893–3901. ACM, 2020. 1, 2

[41] Haoxuan You, Luowei Zhou, Bin Xiao, Noel Codella, Yu Cheng, Ruochen Xu, Shih-Fu Chang, and Lu Yuan. Learning visual representation from modality-shared contrastive language-image pre-training. In *Computer Vision – ECCV 2022, Part XXVII*, pages 69–87. Springer, 2022. 2

[42] Yang Zhao, Chen Zhang, Haifeng Huang, Haoyuan Li, and Zhou Zhao. Towards effective multi-modal interchanges in zero-resource sounding object localization. In *Advances in Neural Information Processing Systems*, pages 38089–38102. Curran Associates, Inc., 2022. 1, 2, 4, 6, 7

[43] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation, 2022. 2, 6, 3, 4, 5

[44] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2023. 1, 2, 6, 3, 4